Journal of Sensor Technology and Application 传感器技术与应用, 2022, 10(2), 128-137
Published Online April 2022 in Hans. http://www.hanspub.org/journal/jsta
https://doi.org/10.12677/jsta.2022.102016

Hans汉斯

基于随机森林算法以及可见－近红外光谱的苹果糖度无损检测

蒋雨鹏，任　玉，蔡红星*，周建伟，王康华，孙　哲

长春理工大学吉林省光谱探测科学与技术高校重点实验室，吉林　长春

收稿日期：2022年3月2日；录用日期：2022年4月2日；发布日期：2022年4月8日

摘　要

本文基于可见－近红外光谱分析技术结合随机森林算法实现不同产地的苹果糖度无损检测。研究通过漫反射采集系统收集三种不同产地苹果的光谱数据后经多种预处理办法比较，采用标准正态变换分别结合偏最小二乘、随机森林算法建立苹果糖度检测通用模型。结果显示该模型预测集相关系数(R_p^2)和预测均方根误差(RMSEP)分别为0.89和0.44，相比偏最小二乘法检测模型相关系数(R_p^2)和预测均方根误差(RMSEP)的0.85和0.47，均有提高。研究扩大了单一品种模型的预测范围，结合随机森林算法有效地提升模型的预测稳健性，对进一步实现水果品质无损检测具有良好的潜在意义。

关键词

光谱学，可溶性固形物，可见－近红外光谱，随机森林，无损检测

Non-Destructive Detection of Apple Sugar Content Based on Random Forest Algorithm and Visible-Near Infrared Spectroscopy

Yupeng Jiang, Yu Ren, Hongxing Cai*, Jianwei Zhou, Kanghua Wang, Zhe Sun

Key Laboratory of Jilin Province Spectral Detection Science and Technology, Changchun University of Science and Technology, Changchun Jilin

Received: Mar. 2nd, 2022; accepted: Apr. 2nd, 2022; published: Apr. 8th, 2022

*通讯作者。

文章引用: 蒋雨鹏, 任玉, 蔡红星, 周建伟, 王康华, 孙哲. 基于随机森林算法以及可见-近红外光谱的苹果糖度无损检测[J]. 传感器技术与应用, 2022, 10(2): 128-137. DOI: 10.12677/jsta.2022.102016

Abstract

This paper is based on visible-near-infrared spectroscopy analysis technology combined with random forest algorithm to achieve non-destructive testing of apple sugar content in different producing areas. The study collects the spectral data of three apples from different origins through the diffuse reflectance collection system and compares them with a variety of preprocessing methods. The standard normal transformation is combined with partial least squares and random forest algorithms to establish a general model for apple sugar content detection. The results show that the correlation coefficient (R_p^2) and root mean square error (RMSEP) of the prediction set of the model are 0.89 and 0.44, respectively. Compared with the partial least square method to detect the correlation coefficient (R_p^2) and root mean square error (RMSEP) of the model 0.85 and 0.47, both improved. The research expanded the prediction range of the single-variety model, combined with the random forest algorithm to effectively improve the prediction robustness of the model, which has good potential significance for the further realization of non-destructive testing of fruit quality.

Keywords

Spectroscopy, Soluble Solids, Visible-Near Infrared Spectroscopy, Random Forest, Non-Destructive Detection

1. 引言

苹果是全球水果产品市场上被广泛选择的水果之一。因富含抗血酸及多酚类化合物等抗氧化成分，可对多种退化性疾病起到预防作用[1]。消费者也同时需要口感和质量更好的水果以满足日益增加的需求。因此，对每个水果进行分类和分级，确保其内部质量很有必要性[2]。苹果的可溶性固形物含量(Soluble Solids Content, SSC)，又称可溶性糖类，涵盖单糖、双糖、多糖[3]，作为主要参数评价苹果的内部品质。传统的破坏性检测方法需要大量的时间和人力，可见－近红外光谱分析技术具备更简单、更快速、更准确的优势被广泛应用于无损评估产品的 SSC 含量，在水果内部质量分类和分级中展现出巨大的潜力[4]。国外对 SSC 含量检测的研究起步时间较早。1998 年，Lammertyn 等人通过 380 nm 到 1650 nm 波长的红外射线完成了苹果的 SSC 含量检测，相干系数达到 0.9 [5]。Salguero-Chaparro L 等(2014)基于近红外光谱技术对橄揽果内部指标实现了在线捡测，表明将近红外光谱检测技术应用在橄揽油加工生产具备可行性[6]。国内方面。2006 年，刘燕德通过近红外光谱对水果糖度及酸度实现无损检测，测试了水果在不同的测试位置、测量距离等因素对模型效果的影响[7]。郭志明(2015)采用自建的苹果在线检测装置，获取苹果的漫反射光谱后通过算法对模型优化开发了准确性较高的短波近红外苹果糖度在线检测装置[8]。

受到果实生长期土壤、天气、光效等不同条件的影响，光学传输特性由于物化特性的不同产生改变[9] [10]。樊书祥等对 4 个不同产地的苹果建立了混合检测模型，得出不同产地通用模型对于提高总体预测模型稳定性方面具有较好的成效[11]。Li X 测定了三产地苹果 SSC 值，预测结果表明，通过原产地判别方法的多产地苹果 SSC 模型可减少原产地对结果的影响[12]。现有研究中，多是探究不同产地或品种

与单一产地的结果对比，对于如何提高不同产地的苹果 SSC 检测通用模型精度及效率报道较少。因此，本研究通过预处理方法减少光谱差异，基于随机森林算法建立非线性模型，提升多产地苹果糖度通用模型预测效果及模型稳定性。研究结果对苹果糖度检测在实际应用中具备生产价值和借鉴意义。

2. 理论部分

2.1. 可见－近红外光谱分析技术

可见－近红外光谱分析技术的研究进展迅速，以非破坏性检测方式为衡量待测品内在品质提供了方法[13]，其原理是具有近红外吸收的物质分子被近红外光通过时会吸收不同的能量，导致不同能级间产生跃迁，通过各谐振子间互相激发，致使近红外区域产生吸收光谱。在该研究中，水果组织内部 O-H，C-H 及 N-H 等化学键倍频与伸缩振动等和近红外光谱吸收特性紧密相关。水果随着自身生长期或贮存期的变化，其中糖含量会随之改变导致组织结构对光吸收或散射的变化。光照射于水果表面时，光被反射、吸收以及散射[14]，散射光随水果内部属性变化的规律被应用于水果的内部品质无损检测研究。

2.2. 漫反射原理

当一束漫反射光进入水果内部会发生吸收以及多次反射、折射等现象，返回而出的光承载了样品的光谱信息。相比其他采集方式，漫反射光更能反映水果的组织结构特点和内部成分特性。因此，漫反射采集方式被广泛应用在大部分水果内部品质无损检测研究中[15]。

Kubelka-Munk 函数为漫反射检测方法的光谱进行定量分析，其表达式为：

$$R'_{\infty} = 1 + \frac{K}{S} - \sqrt{\left(\frac{K}{S}\right)^2 + 2\left(\frac{K}{S}\right)} \tag{1}$$

上述表达式中，漫反射体的内部化学组分通过漫反射体吸收系数 K 表示；散射系数表示为 S，通过漫反射体的外部物理特征确定。K/S 的函数用来表达待测物的绝对漫反射率，表示出射光与入射光的比值。

因此相对漫反射率为：

$$R_{\infty} = \frac{R'_{\infty品}}{R'_{\infty比}} = 1 + \frac{K}{S} - \sqrt{\left(\frac{K}{S}\right)^2 + 2\left(\frac{K}{S}\right)} \tag{2}$$

漫反射分析中，上式的 R_{∞} 与样品的组分浓度并不构成线性关系，与其构成线性关系的函数为反射吸光度[16]。

漫反射吸光度 A 表达式为：

$$A = \log\left[\frac{1}{R_{\infty}}\right] = -\log\left[1 + \frac{K}{S} - \sqrt{\left(\frac{K}{S}\right)^2 + 2\left(\frac{K}{S}\right)}\right] \tag{3}$$

通过可见－近红外漫反射光谱转化为吸光度可与物质的成份或内部性质建立关联，从而建立相应的关联模型。

2.3. 随机森林算法

随机森林算法(Random Forest)在机器学习中成为被广泛选择且应用的算法之一。其以简单、灵活的特点应用于分类和回归任务。它构建的“森林”是决策树的集合。决策树的形状为树状结构，其中内部的一个节点表达一个测试属性；一个分支表达一个测试输出；一个叶节点则表达一个类别。随机森林通过集成学习的思想，以决策树为基本单元，在构建过程中遵循样本抽样随机有放回和特征随机对预测结

果进行投票，最后根据决策树模型的平均值来获取最终结果。判断随机森林模型结果优劣的指标主要受到训练样本个数、样本变量、决策树预测能力以及各决策树之间相关性的影响[17]。相关系数(R^2)、校正均方根误差(RMSEC)和预测均方根误差(RMSEP)作为模型的评价指标，决定了模型效果优劣。

决定系数：

$$R_c^2, R_p^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{i,actual} - y_{i,predicted}\right)}{\sum_{i=1}^{n}\left(y_{i,actual} - \overline{y_{i,actual}}\right)^2} \tag{4}$$

校正均方根误差：

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{i}^{n}\left(y_{i,actual} - y_{i,predicted}\right)^2}{n-1}} \tag{5}$$

预测均方根误差：

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i}^{n}\left(y_{i,actual} - y_{i,predicted}\right)^2}{n-1}} \tag{6}$$

校正集相关系数、均方根误差用来评价水果糖度和光谱数据之间的校正关系；预测集相关系数、均方根误差可以判断模型的预测效果。当模型的相关系数越接近 1 时，证明其自变量对因变量的解释程度越高，模型拟合效果越佳；校正均方根误差越小则证明模型回归性强，样本相关性紧密；预测均方根越小，则表明模型预测能力强，二者越相近，模型的泛化能力越稳固。

3. 实验部分

3.1. 实验样品

实验样品为同一日期购自相同超市的山东富士、陕西富士、新疆阿克苏各 12 个，共计 36 个样品。人工挑选体型相近、表面无明显伤痕瑕疵作为样品。使用毛巾对每个样品表面进行擦拭清洁。每个水果根据种类依次进行编号标记，在垂直于茎轴的赤道位置每间隔 120˚标记取样点。实验前将所有样品在实验室静置 24 H，使其接近实验室室温、湿度，做等温处理。

3.2. 实验系统搭建

基于漫反射采集原理搭建了试验检测系统，如图 1 所示为漫反射检测系统图。采用溴钨灯作为试验光源，光源出光口与光纤探头紧贴样品表面且相距 3 cm。入射光通过样品表面，将光信息传递到实验水果样本的表面上，漫反射出射光将携载着实验对象样本的内部结构和信息传递回光纤探头；最终，漫反射光承载光谱信息，传递到光谱仪后的光谱数据保存在实验计算机中。

采集前，实验设备需要预热 15 min，保证平稳运行。采用 SpectraSuite 光谱采集软件，积分时间设置 1 s，平均次数为 3，对每个采集点依次进行光谱采集，共采集 108 组光谱数据。

3.3. 理化值测定

光谱采集后，通过糖度折光仪实现样品的糖度值测定。糖度计根据光的折射原理可以准确地测出溶液中的含糖量。具体的操作方法是：穿戴一次性手套，在之前标记的光谱采集点，用刀取等量苹果果肉，并用手压捏出果汁，滴在折光仪的棱镜表面中央约 2~3 滴，迅速合上棱镜盖后静待 10 秒钟，调节目镜使刻度线清晰并读取实际糖分值。对每个样品的采集点重复采集读数三次，取平均值后作为对应测量样品的真实糖度值。

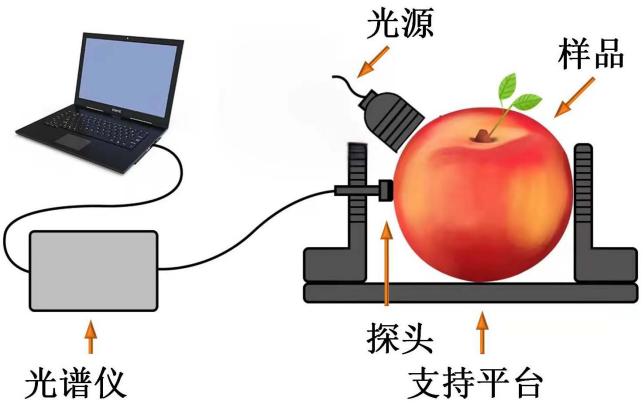


Figure 1. Diffuse reflection experimental system diagram
图 1. 漫反射实验系统图

3.4. 光谱预处理方法

在对漫反射光谱的分析过程中，经常会受到样品、仪器或其它因素的影响，出现基线倾移之类的现象。研究分别采用一阶导数、多元散射校正、S-G 卷积平滑、标准正态变换四种方法作为光谱数据的预处理方式，优选出最佳预处理方法应用于后续模型建立。

4. 结果与讨论

4.1. 样本理化值统计

样本集 SSC 结果统计见下表 1，其中 SSC 的范围在 10.2˚Brix~15.2˚Brix 之间，标准差为 1.01˚Brix。样本范围较大，有利于构建模型。

Table 1. System resulting data of standard experiment
表 1. 标准试验系统结果数据

产地	数量	范围	平均值	标准差
山东富士	36	10.2~14.2	12.88	1.06
陕西富士	36	11.0~15.0	12.25	0.90
阿克苏	36	12.4~15.2	13.75	0.82
总计	108	10.2~15.2	12.96	1.01

4.2. 样本漫反射光谱响应特征

由于光谱两端主要为无效信息且包含较大噪声。后续研究选择 650~900 nm 波段进行探究分析。样本光谱图如图 2 所示。三种苹果的光谱曲线有很明显的吸收特性且总体趋势相似性较高。其中造成波长 675 nm 处光谱吸光度强度变化的原因可能为果肉细胞中叶绿素和类胡萝卜素对光谱的吸收[18]，与样本的表皮颜色差异和不同成熟期样品内部的成分差异相关。740 附近的吸收峰可能与 O-H 键的三级倍频和 C-H 键的四级倍频伸缩振动存在关联[19]。近红外波长下的光谱吸收大多数与 C-H 和 O-H 化学键的吸收相关，而这些化学键又是组成水分、可溶性糖、纤维素和果胶等物质的基础形式。840 nm 附近的较弱波峰与 N-H 三级倍频伸缩振动相关。它们作为基础化学键构成有机化学物。

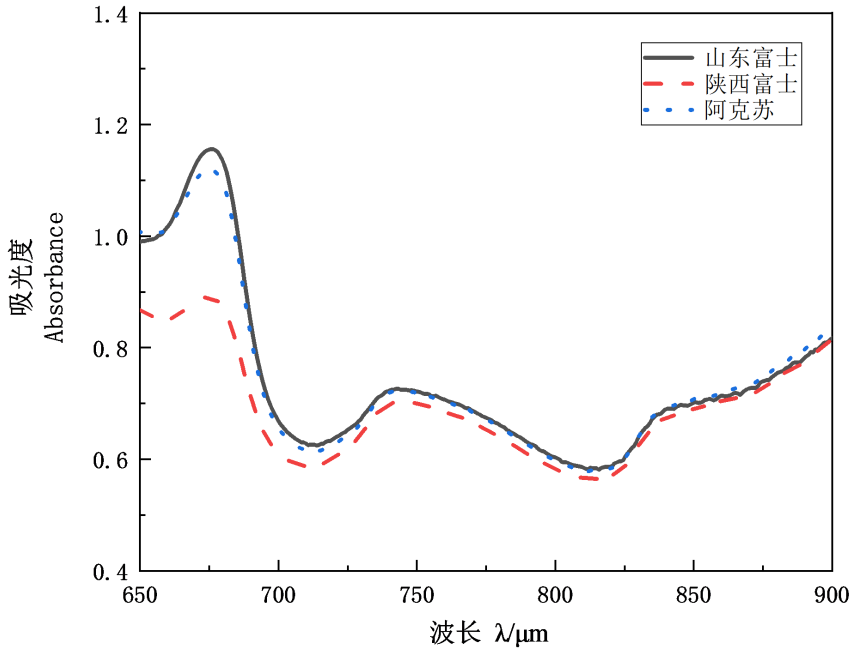


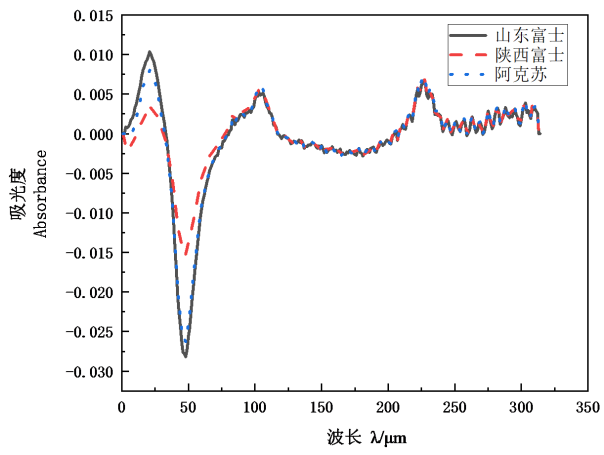
Figure 2. Mean sample spectrum
图 2. 样品平均光谱

4.3. 样品光谱预处理

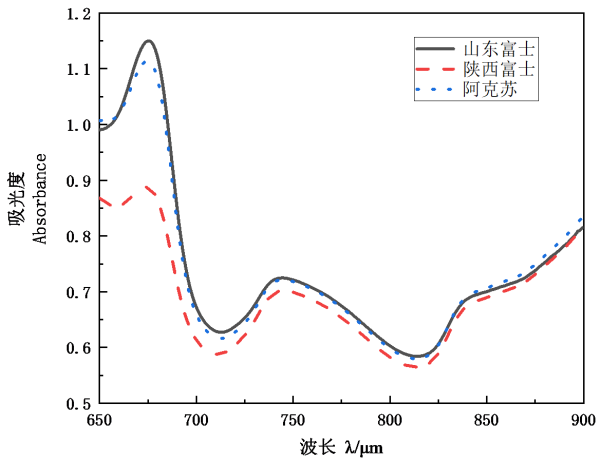
试验选用了一阶导数(1st)、S-G 卷积平滑、标准正态变换(SNV)及多元散射校正(MSC)四种方法分别对原始光谱数据进行预处理，从而消除光谱数据中基线和噪声等因素对光谱的扰乱，获取信噪比较高的光谱数据，对于提高预测模型的稳定性具有很大的帮助作用。苹果样本经过 1st (a)、S-G 卷积平滑(b)、SNV (c)和 MSC (d)预处理后的平均光谱吸光度曲线如图 3。

4.4. 模型建立

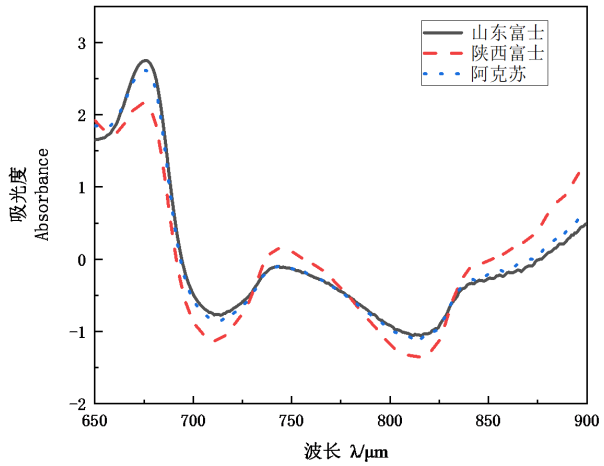
在建立检测模型前，采用随机法对实验样本按 2:1 的比例分成校正集和预测集。其中校正集用来建造校正模型，预测集评定校正模型的决定参数。



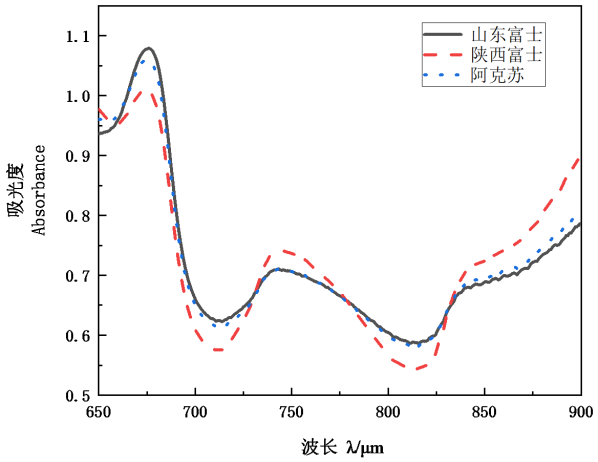
(a) 一阶导数平均光谱



(b) S-G 卷积平滑平均光谱



(c) 标准正态变换平均光谱



(d) 多元散射校正平均光谱

Figure 3. The average spectrum of the sample after pretreatment
图 3. 经过预处理后的样品平均光谱

4.4.1. 偏最小二乘法建模

采用原始光谱数据与分别经过 4 种预处理方法后的光谱数据分别结合偏最小二乘法(PLS)建模。最终结果如表 2 所示。

Table 2. PLS model evaluation index
表 2. PLS 模型评价指标

预处理	校证集		预测集	
	R_c^2	RMSEC	R_p^2	RMSEP
未处理	0.79	0.53	0.78	0.57
1st	0.79	0.55	0.79	0.59
S-G	0.80	0.51	0.79	0.55
SNV	0.87	0.42	0.85	0.47
MSC	0.73	0.64	0.71	0.67

由表 2 可知采用 PLS 结合不同预处理方法对苹果通用糖度检测模型的评价指标。观察得出采用一阶求导法与未经预处理光谱的模型评价指标较为接近；经 S-G 卷积平滑的光谱模型指标略优于未处理光谱检测模型；采用多元散射法建立的光谱模型相比下效果不佳。比较可知，采用标准正态变换结合 PLS 检测模型效果最好。因此，本次试验经比较后采用标准正态变换(SNV)用于 PLS 建模分析及后续的随机森林算法建模分析。图 4 为 SNV-PLS 预测模型中苹果 SSC 值的预测散点图，可以看出，预测值集中在目标线附近，预测效果较好。

4.4.2. 随机森林法建模

通过标准正态变换(SNV)处理后的光谱结合随机森林算法建立苹果糖度检测通用模型。与 PLS 模型采用相同的样本数据。表 3 为采用 SNV 预处理结合随机森林算法苹果通用模型的评价指标。图 5 中，预测模型的 SSC 预测值散点集中于目标线附近，偏差很小。

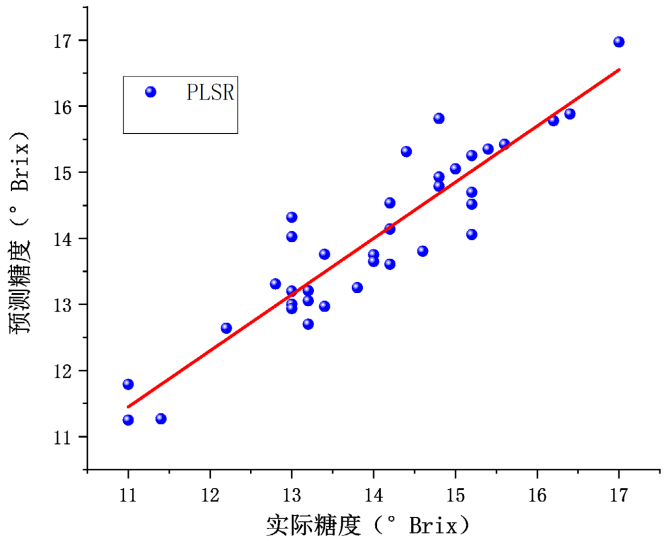


Figure 4. PLS model prediction set sample results
图 4. PLS 模型预测集样本结果

Table 3. RF model evaluation index
表 3. RF 模型评价指标

SNV	校正集		预测集	
	R_c^2	RMSEC	R_p^2	RMSEP
RF	0.91	0.41	0.89	0.44

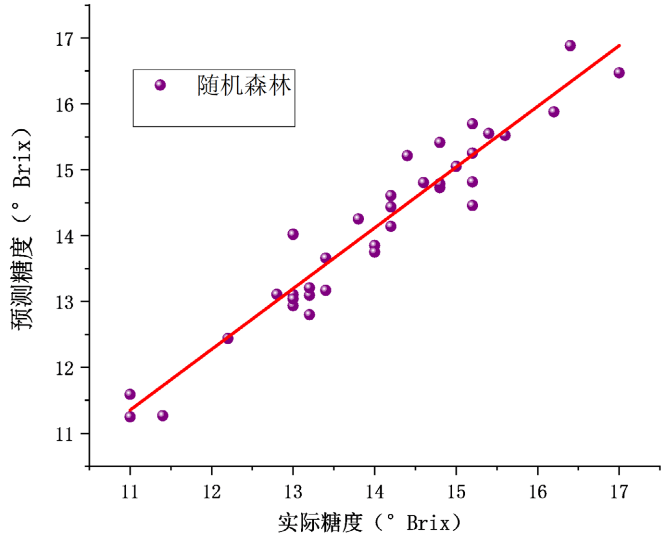


Figure 5. RF model prediction set sample results
图 5. RF 模型预测集样本结果

5. 结论

本研究基于可见－近红外光谱分析技术，利用随机森林算法，通过漫反射采集方式采集了三种不同产地苹果的光谱数据，经过四种光谱预处理方法分别结合 PLS 建立糖度检测模型比较后，采用效果最佳

的标准正态变换(SNV)结合随机森林回归算法建立了不同产地苹果的糖度检测通用模型，并与利用偏最小二乘算法的糖度通用模型做出比较。实验结果表示，采用随机森林算法构建的通用模型校正集的相关系数与校正均方根误差分别为 0.91 和 0.41，预测集相关系数和预测均方根误差分别为 0.89 和 0.44；结合偏最小二乘法建立的通用模型其校正集相关系数与校正均方根误差为 0.87 和 0.42，预测集相关系数、预测均方根误差分别为 0.85、0.47。结论表明，采用 SNV 预处理方法结合随机森林算法建立的通用检测模型在预测不同产地苹果糖度时，预测精度相对偏最小二乘法有较大提升，有效减小了由产地不同引起的光谱差异。同时表明预测精度在受到生物多样性变化时有较好的适应性及准确性，使得预测模型面对未知变化更加稳定。实验结果表明随机森林算法通用模型对 3 种产地苹果糖度具有出色的预测能力，试验结论与研究方法对今后的水果糖度检测通用模型研究具有参考价值，可缩减不同产地苹果糖度检测模型建造过程中的模型维护成本，为水果采摘后的分选等商品化处理具有参考价值。

基金项目

吉林省自然科学基金项目(20200201257JC, 2020)。

参考文献

[1] Mendoza, F., Lu, R. and Cen, H. (2014) Grading of Apples Based on Firmness and Soluble Solids Content Using VIS/SWNIR Spectroscopy and Spectral Scattering Techniques. *Journal of Food Engineering*, **125**, 59-68. https://doi.org/10.1016/j.jfoodeng.2013.10.022

[2] Beghi, R., Giovanelli, G., Malegori, C., *et al*. (2014) Testing of a VIS-NIR System for the Monitoring of Long-Term Apple Storage. *Food & Bioprocess Technology*, **7**, 2134-2143. https://doi.org/10.1007/s11947-014-1294-x

[3] Mendoza, F., Lu, R., Ariana, D., *et al*. (2011) Integrated Spectral and Image Analysis of Hyperspectral Scattering Data for Prediction of Apple Fruit Firmness and Soluble Solids Content. *Postharvest Biology & Technology*, **62**, 149-160. https://doi.org/10.1016/j.postharvbio.2011.05.009

[4] Temma, T., Hanamatsu, K. and Shinoki, F. (2002) Measuring the Sugar Content of Apples and Apple Juice by Near Infrared Spectroscopy. *Optical Review*, **9**, 40-44. https://doi.org/10.1007/s10043-002-0040-1

[5] Lammertyn, J., Nicola, B., Ooms, K., *et al*. (2001) Non-Destructive Measurement of Acidity, Soluble Solids, and Firmness of Jonagold Apples Using Nir-Spectroscopy. *Transactions of the ASAE*, **41**, 1089-1094. https://doi.org/10.13031/2013.17238

[6] Salguero-Chaparro, L. and Rodríguez, F. (2014) On-Line versus Off-Line NIRS Analysis of Intact Olives. *LWT—Food Science and Technology*, **56**, 363-369. https://doi.org/10.1016/j.lwt.2013.11.032

[7] 刘燕德. 水果糖度和酸度的近红外光谱无损检测研究[D]: [博士学位论文]. 杭州: 浙江大学, 2006.

[8] 郭志明. 基于近红外光谱及成像的苹果品质无损检测方法和装置研究[D]: [博士学位论文]. 北京: 中国农业大学, 2015.

[9] 刘燕德, 徐海, 孙旭东, 等. 不同产地苹果糖度可见近红外光谱在线检测[J]. 中国光学, 2020, 13(3): 482-491.

[10] Zhang, B., Dai, D., Huang, J., *et al*. (2017) Influence of Physical and Biological Variability and Solution Methods in Fruit and Vegetable Quality Non-Destructive Inspection by Using Imaging and Near-Infrared Spectroscopy Techniques: A Review. *Critical Reviews in Food Science & Nutrition*, **58**, 2099-2118. https://doi.org/10.1080/10408398.2017.1300789

[11] 樊书祥, 黄文倩, 郭志明, 等. 苹果产地差异对可溶性固形物近红外光谱检测模型影响的研究[J]. 分析化学, 2015, 43(2): 239-244.

[12] Li, X., Huang, J., Xiong, Y., *et al*. (2018) Determination of Soluble Solid Content in Multi-Origin "Fuji" Apples by Using FT-NIR Spectroscopy and an Origin Discriminant Strategy. *Computers and Electronics in Agriculture*, **155**, 23-31. https://doi.org/10.1016/j.compag.2018.10.003

[13] Travers, S., Bertelsen, M.G., Petersen, K.K., *et al*. (2014) Predicting Pear (cv. Clara Frijs) Dry Matter and Soluble Solids Content with Near Infrared Spectroscopy. *LWT—Food Science and Technology*, **59**, 1107-1113. https://doi.org/10.1016/j.lwt.2014.04.048

[14] 李晓红, 张想汉, 唐春晓, 等. 苹果组织中散射光场分布的光学特性研究[J]. 食品安全质量检测学报, 2014(4): 1166-1172.

[15] 邱雁. 漫反射光谱的理论与应用研究[D]: [硕士学位论文]. 上海: 同济大学, 2007.

[16] 谭保华, 肖腾飞, 刘琼磊, 等. 典型经济水果近红外漫反射无损检测及其光谱数据分析[J]. 湖北农业科学, 2020, 59(12): 154-158.

[17] 李盛芳. 基于机器学习的水果糖分近红外光谱检测方法研究[D]: [硕士学位论文]. 太原: 太原理工大学.

[18] Jamshidi, B., Minaei, S., Mohajerani, E., *et al*. (2012) Reflectance Vis/NIR Spectroscopy for Nondestructive Taste Characterization of Valencia Oranges. *Computers & Electronics in Agriculture*, **85**, 64-69. https://doi.org/10.1016/j.compag.2012.03.008

[19] Yuan, L.-M., *et al*. (2016) Nondestructive Measurement of Soluble Solids Content in Apples by a Portable Fruit Analyzer. *Food Analytical Methods*, **9**, 785-794. https://doi.org/10.1007/s12161-015-0251-2